PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
H04Q 11/04

A2

(11) International Publication Number: **WO 99/25147**

(43) International Publication Date: 20 May 1999 (20.05.99)

(21) International Application Number: PCT/FI98/00872

(22) International Filing Date: 10 November 1998 (10.11.98)

(30) Priority Data:
974216 12 November 1997 (12.11.97) FI

(71) Applicant *(for all designated States except US)*: NOKIA TELECOMMUNICATIONS OY [FI/FI]; Keilalahdentie 4, FIN–02150 Espoo (FI).

(72) Inventor; and
(75) Inventor/Applicant *(for US only)*: MA, Jian [FI/FI]; Pihlajamarjantie 13–15 C1, FIN–02940 Espoo (FI).

(74) Agent: PATENT AGENCY COMPATENT LTD.; Teollisuuskatu 33, P.O. Box 156, FIN–00511 Helsinki (FI).

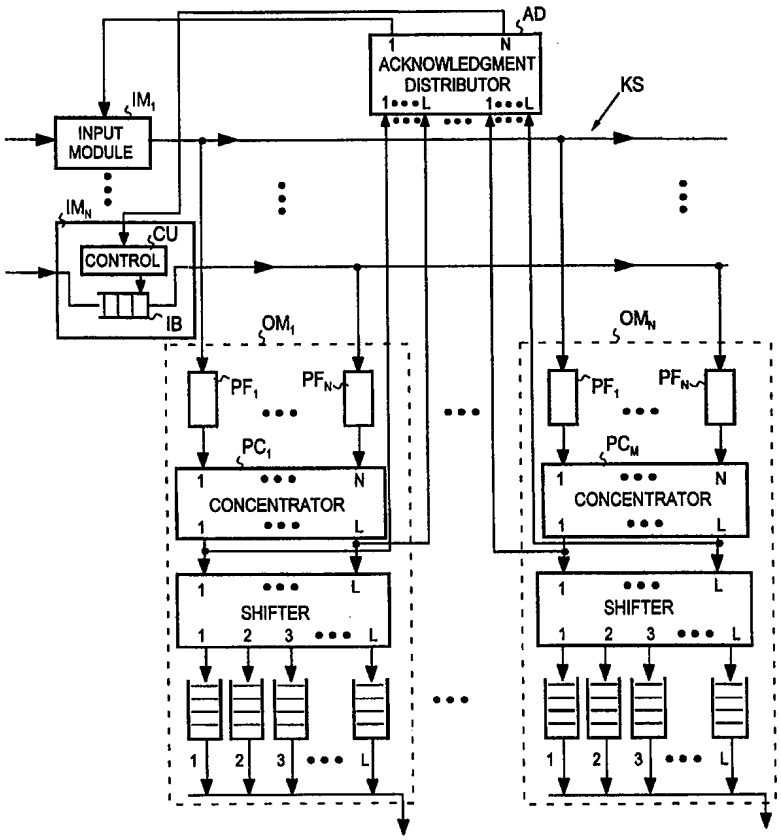
(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
Without international search report and to be republished upon receipt of that report.

(54) Title: A FRAME DISCARD MECHANISM FOR PACKET SWITCHES

(57) Abstract

The invention relates to a frame discard method for a packet switch and to a packet switch. The method comprises the steps of (a) receiving packets belonging to at least one transmission connection, consecutive packets of an individual transmission connection forming frames, (b) switching packets from N input ports of the switch to N output ports of the switch, through at least one intermediate port, and (c) discarding complete frames, if the load level in the switch exceeds a predetermined threshold. In order to combine a good frame level performance with implementation simplicity, the packets are discarded in a switch where the maximum number of packets that can be transmitted simultaneously to one port is smaller than N, in such a way that (i) when the number of packets simultaneously contending for an individual port exceeds said maximum number, at least one first packet of a frame is chosen to be discarded from among said packets, and (ii) once a first packet of a frame has been discarded, the remaining packets of the same frame are discarded, regardless of the current buffering capacity of the switch.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

A frame discard mechanism for packet switches

Field of the invention

This invention relates generally to a packet discard mechanism for packet switches. More particularly, the invention relates to a packet discard mechanism which discards whole frames instead of randomly discarding individual packets (or cells).

Background of the invention

In a packet switched network, most of the higher level data units to be delivered through the switches of the network are too large to be transferred in a single packet and must therefore be segmented for delivery. In this context, these higher level data units, which are used in various kinds of applications, are called frames. Figure 1 illustrates the relationship between packets and frames as used in this context; a transmission frame consists of M consecutive packets, each having a fixed length. In a TCP/IP over ATM network, for example, a frame corresponds to an IP datagram and the packets correspond to ATM cells. The frame size (the number of packets in a frame) depends on the type of application using a transmission connection.

If one or more of the packets of a frame are missing at the destination, the frame cannot be reassembled and must be discarded. Therefore, if the performance of the network is evaluated from the point of view of the applications sending and receiving these higher level data units, it should be clear that the end-to-end delay and the loss rate of these data units (i.e. frames) are more relevant performance indicators than the end-to-end delay and the loss rate of individual data packets.

When random packet dropping policies are used in the network, it is more likely that the dropped packets belong to different frames than to the same frame. Therefore, the packet loss rate of such drop mechanisms cannot be an indication of the quality of service (QoS) on the application level.

To eliminate this drawback, more sophisticated loss mechanisms have been proposed. One is what is called the early packet discard (EPD) scheme, which discards whole frames instead of randomly discarding packets. The simplest way of implementing EPD is to set a threshold value in each buffer of the switch. The first packet of any incoming frame is discarded when the fill rate of the buffer exceeds the threshold value. Once the first packet of a

frame is discarded, the remaining packets of the frame are also discarded, even if the fill rate has decreased to below the threshold. However, a packet is not discarded if the first packet of the same frame has been accepted, unless the entire buffer is full. The threshold value should be chosen in such a way that, on the one hand, no buffer overflow occurs and, on the other hand, no corrupted frames are transmitted.

For those interested in the subject, a detailed description of the EPD method can be found, for example, in an article by A. Romanow and S. Floyd, *Dynamics of TCP Traffic over ATM Networks*, Proc. ACM SIGCOMM '94, pp. 79-88, August 1994.

One drawback of the EPD method is that it deals unfairly with the different users. This is due to the fact that the EPD scheme discards complete frames from all connections, without taking into account their current rates or their relative shares in the buffer, i.e. without taking into account their relative contribution to an overload situation. To remedy this drawback, several variations for selective drop policies have been proposed.

Nevertheless, the EPD method and its variations still have the disadvantage of being restricted for use only in association with the buffers of the switch, the method becoming active only when the fill rate of a buffer exceeds the threshold.

The EPD method and its variations have been studied assuming that the switch is an output-buffered full speed switch, i.e. that the switch capacity equals to the dimension (number of output/input ports) of the switch. A further drawback relating to this type of switch is that the internal switching speed of the switch becomes high if the dimension of the switch is large. This in turn means that powerful (and expensive) chips must be used.

Summary of the invention

The purpose of the invention is to eliminate the above-mentioned drawbacks and to create a new method which makes it possible to obtain a frame level performance equal to that of the above prior art methods by means of a less complex and less expensive switch structure, and without the need to control the fill rates of individual buffers.

This goal can be attained by using the solution defined in the independent patent claims.

According to the invention, a frame drop mechanism is introduced into a switch having a capacity smaller, preferably essentially smaller than the number of inputs and outputs of the switch. This is done so that the first packet of one or more frames is discarded if the number of packets simultaneously destined for the same port of the switch exceeds a predetermined value (the switch capacity). Once the first packet of a frame has been discarded, the remaining packets of the same frame are discarded when they access the switch, regardless of the current buffering capacity of the switch, i.e. the other packets of corrupted frames are prevented from entering the switching fabric.

In other words, the idea of the invention is to introduce a drop mechanism into the switching fabric so that the variable which controls the activation and deactivation of the drop mechanism indicates the number of packets simultaneously contending for the same port. This port can be an output port of the whole switch or any output port within the switching fabric. As the feature of several packets contending simultaneously for the same port relates only to output-buffered switches, the switch according to the invention must be an output-buffered switch.

By means of the solution according to the present invention, the implementation simplicity generally characteristic of input-queued switches, the good throughput performance generally characteristic of output-queued switches, and a low frame loss rate can be combined in a single switch more efficiently than before.

Implementation simplicity can be improved, as the internal switching speed of the switch according to the invention is not proportional to the dimension of the switch but to the capacity of the switch, which is smaller, preferably much smaller, than the dimension.

The switch according to the invention is preferably based on a knockout type structure, as this kind of structure can easily be modified to fulfill the functionality required by the invention.

Brief description of the drawings

In the following, the invention and its preferred embodiments are described in closer detail with reference to examples shown in the appended drawings, wherein

Figure 1 illustrates the relationship between packets and frames as used here,

Figure 2 shows the general architecture of a Knockout switch,

Figure 3 is a block diagram of a single output module of a Knockout switch,

Figure 4 is a flow chart illustrating the steps of the method according to the invention,

Figure 5 shows a preferred embodiment of the switch according to the present invention,

Figures 6a to 6d illustrate the states of an individual switching element of a prioritized Knockout concentrator,

Figure 7 illustrates the classification of packets in the switch of Figure 5,

Figure 8 illustrates one possible implementation of the acknowledgment distributor of Figure 5,

Figure 9a...9d illustrate the frame drop process in the switch according to the invention,

Figure 10 illustrates a second preferred embodiment of the switch according to the invention,

Figure 11 is a block diagram of an individual packet filter of Figure 10,

Figure 12 is a flow chart illustrating the functions of an individual packet filter of Figure 10, and

Figure 13 illustrates the routing of acknowledgments in the output module of Figure 10.

Detailed description of the invention

As indicated above, the method according to the present invention is used in an output-buffered switch which switches fixed-length packets from N inputs to N outputs. In other words, the switch can be any packet switch suitable for switching fixed-length packets, such as an ATM switch. Further, the switch capacity L, which is defined as the maximum number of packets that can be transmitted simultaneously to one output port in one time slot, is smaller than N, i.e. $L<N$. A widely-known switch architecture fulfilling these prerequisites is the Knockout architecture. Therefore, the invention is described below by using the Knockout switch as an exemplary switch structure to which the invention is applied.

As known, the design of the Knockout switch is based on the knockout principle: if the arrivals of packets on N different input lines are statis-

tically independent, the probability is extremely low that more than a handful of simultaneous packets, i.e. L packets (L<<N), are destined for any particular output port, even for arbitrarily large dimensions (N×N) of the switch.

The basic structure of the Knockout switch is illustrated in Figure 2. A packet arriving at one of N inputs is placed onto a broadcast bus from which it is transferred to each of the N output modules OM_i (i=1,2,..N). The output modules read the header of the packets, accepting those packets destined for that output and buffering those packets which cannot be immediately placed onto the output link, i.e. when two or more packets are contending simultaneously for the same output.

A block diagram of an output module OM_i is shown in Figure 3. Each of the input buses is connected to a packet filter PF_i (i=1,2,..N), i.e. there are N packet filters at the input of each output module. Each packet filter reads the header of the arriving packet and delivers it to a concentrator CC if the header indicates that the packet is intended for that output. Otherwise, the packet is ignored. Consequently, packets that appear at the outputs of the packet filters of a certain output module belong to the switch output being served by that particular output module.

The concentrator comprises N inputs but only L (L<N) outputs. In any time slot, the concentrator finds all active packets appearing at its inputs. If L or fewer packets are found, these packets are placed in the concentrator outputs, with the left-most outputs being filled first. On the other hand, if more than L packets arrive simultaneously (in the same time slot) at the concentrator, all but L of those packets will be lost. By properly choosing the value of L, the rate of loss resulting from this "built-in" packet loss mechanism can be maintained at an acceptable level. Thus, L represents the maximum number of packets that can be simultaneously transferred to a particular output port, i.e. L represents the switch capacity.

The outputs of the concentrator are connected to a shifter SH having L inputs and L outputs. As the concentrator fills its left-most outputs first, the left-most buffers would tend to fill up if the packets were connected directly from the concentrator to the L output buffers. The purpose of the shifter is to make sure that the L output buffers OB_i (i=1,2,...L) are filled as evenly as possible.

The output line OL_i connected to the outputs of the output buffers fetches packets cyclically from the output buffers; in time slot 1 the output line

fetches a packet from buffer OB_1, in time slot 2 a packet from buffer OB_2, and so on. After the output line has fetched a packet from the last (L^{th}) buffer, the next packet is again fetched from the first buffer (OB_1).

As the Knockout switch is widely-known, an interested reader can find detailed descriptions from the many articles or books describing its architecture and functionality. The Knockout switch is also described in U.S. patent 4,760,570. Therefore, the features of the Knockout switch are not described in more detail here. Instead, the following description describes the modifications introduced to the Knockout architecture by the present invention.

According to the present invention, whole frames are discarded in the switch, instead of dropping packets randomly as is done in the Knockout switch. The frames are discarded in such a way that the switch begins to drop the first packet of one or more frames randomly, if the number of packets simultaneously destined for an output port exceeds the switch capacity L. The switch identifies the frames to which the discarded packets belong and continues to discard the other packets from those frames. This functionality is described in more detail in the following.

Whenever the number of incoming packets destined simultaneously for the same output port exceeds the switch capacity L, the switch begins to discard (knock out) the first packets of some of the incoming frames. Once a first packet of a certain frame is discarded (knocked out), the remaining packets of the same frame are also discarded when they access the switch, even if the switch would have enough buffering capacity to accept said packets. However, a packet is never discarded if the first packet of the same frame has already been accepted by the output port of the switch, unless the entire buffer in question is full.

To prevent the packets of corrupted frames from entering the switching fabric, each input port of the switch checks whether the packet leaving for the switch belongs to a corrupted frame, i.e. whether the first packet of the frame which the packet belongs to was discarded earlier in the switch fabric. Therefore, there are no packets entering the switching fabric that belong to a corrupted frame.

Figure 4 is a flow chart illustrating the basic principle of the invention. The switch continuously discards incoming packets which belong to corrupted frames (phases 41 and 42). Should the number K of packets simultaneously destined for an output port exceed the switch capacity L, the switch

discards a certain number of first packets of frames (i.e. (K-L) packets) by choosing the first packets to be discarded randomly (phases 43 and 44). In other words, the (K-L) frames to be discarded are chosen randomly. As explained below, there are always at least (K-L) first packets in competition for a port when K is greater than L. In Figure 4, the drop process is divided into two consecutive steps, as these steps are performed in different parts of the switch.

In order to be able to check whether an incoming packet belongs to a corrupted frame, the input port of the switch must know if a previous packet of the same frame was accepted or discarded in the switching fabric. In other words, acknowledgments must be sent from the switching fabric to the input ports.

The above principle can be implemented in several ways. In the following, two preferred embodiments based on the Knockout switch are shown in more detail.

Figure 5 shows the first embodiment, comprising a Knockout switch KS, N input modules IM_i (i=1,2,...N), one for each input bus of the Knockout switch, and an acknowledgment distributor AD. In this embodiment the input modules perform step 1 of Figure 4, whereas the concentrators perform step 2.

The knockout switch KS of Figure 5 is otherwise similar to the commonly known Knockout switch, except that the concentrators PC_i (i=1,2,...N) drop first packets of frames, instead of dropping packets randomly. To achieve this, what are known as prioritized concentrators must be used, and the incoming packets must be divided into two classes in terms of loss probability, i.e. a high priority class and a low priority class. The division is done so that the incoming first packet of each frame is marked as a low priority packet and the remaining packets (i.e. those which are not the first packet in a frame) are marked as high priority packets (if not discarded before that).

The internal structure of a prioritized concentrator is similar to that of a common Knockout concentrator, except that the internal 2×2 switching elements function so that a high priority packet always wins the competition against a low priority packet. Therefore, a low priority packet is passed to the next level of competition only when no high priority packet is present at the inputs of a 2×2 switching element. If two high-priority packets are present at the

inputs of a 2×2 switching element, the packet at the right input loses the competition.

Figures 6a to 6d illustrate the states of an individual switching element SE of a prioritized concentrator PC_i, assuming that a packet is present at both inputs.

The classification of the packets is illustrated in Figure 7, where Kh represents the number of high-priority packets and Kl the number of low-priority packets destined for a particular output in a single time slot. In the prioritized concentrators PC_i, (K-L) low priority packets are dropped if K>L.

It is to be noted that high priority packets will never be dropped in the concentrators because the number of contending high priority packets will never exceed L. This is because all other packets (which would be marked as high priority packets) of corrupted frames are not allowed to enter the switching fabric. This is also why there are always at least (K-L) first packets in the competition for an output if K is greater than L. This can be proved as follows:

If K>L, there are two alternatives depending on the value of Kh,

(1) let us assume that Kh>L. This would mean that the number of packets accepted by the switching fabric in the previous time slot is greater than the switch capacity (as one output port of the switch can only accept up to L simultaneous packets). Therefore, this assumption is false.

(2) let us assume that $Kh \leq L$. Then $Kl = K-Kh \geq K-L$, i.e. there are at least (K-L) low priority packets (first packets) in competition for the output.

The acknowledgment distributor AD, which is common to all output modules, has N×L inputs and N outputs. The first L inputs are connected to the outputs of the concentrator of the first output module, the next L inputs to the outputs of the concentrator of the second output module, etc. If a packet passes through a concentrator PC_i, an acknowledgment is generated and sent through the distributor to the correct input module.

The acknowledgment distributor AD can be, for example, a simple crossbar matrix, as shown in Figure 8. There is only one connection in each row and column in the crossbar matrix in each time slot. Figure 8 shows how an acknowledgment is transferred if a packet from input module IM_2 is accepted by output module OM_1. It is assumed that the packet will emerge from the output L of the concentrator of said output module. An acknowledgment message will then be generated at said output, the message including, in addition to other data, the address of the input port from which the packet arrived

(i.e. source address). On the basis of this source address, the switching element (marked with a circle in Figure 8) corresponding to the source address connects the message to the correct input module. It is thus to be noted that the self-routing tag attached to the packets (in the input modules) includes the addresses of both the input and the output ports.

In each time slot, each input module IM_i (i=1,2,...N) checks whether the packet leaving for the switch is the first packet of a frame. If this is the case, the input module marks the packet as a low priority packet and passes it to the switching fabric. If the packet is not a first packet of a frame, the input module checks, using the received acknowledgments, whether the previous packet belonging to the same connection was discarded in the switch. If the previous packet was discarded, the input module discards the packet, but if the previous packet was accepted, the input module marks the packet as a high-priority packet and passes it to the switching fabric.

As shown in association with input module IM_N in Figure 5, the above mentioned functionality can be implemented, for example, by means of a separate control unit CU which receives the acknowledgments, reads the header of the packet which is at the tip of the input buffer IB, and marks or discards said packet.

Figures 9a...9d illustrate an example of the frame drop process in the switch according to the invention. In this example, the switch has four inputs and four outputs, and the switch capacity L equals two. Low priority packets are shown as hatched, and the numbers inside the packets indicate the output port which is the destination of the packet.

In time slot 1 (Figure 9a), the two packets destined for output 1 are accepted because L=2. In time slot 2 (Figure 9b), three packets are heading for output 1, so one of them must be discarded. In this case, the packet from input 2 is discarded in the concentrator, because it is the only low priority packet (first packet of a frame). The two other packets are accepted because they are not the first packet of a frame and because the preceding packets in the same frames were accepted. In time slot 3 (Figure 9c), the packet from input 2 is dropped at the input port 2 because the previous packet of the frame was dropped. In time slot 4 (Figure 9d), the packet from input 2 is again dropped, for the same reason. Nevertheless, there are still three packets destined for output 1. The packet from input 1 is accepted because it is a high pri-

ority packet, whereas one of the two first packets contending for output 1 is randomly dropped.

In the embodiment of Figure 5, the switch utilizes a large centralized acknowledgment distributor. Alternatively, the drop mechanism can be performed in the output modules only. This embodiment is illustrated in Figure 10, which shows one output module of such a switch. The overall structure of the switch is that shown in Figure 2. As shown in Figure 10, each output module comprises N packet filters PF'_i (i=1,2,...,N), a prioritized concentrator PC_i with N inputs and L outputs, a shifter, L output buffers, and an acknowledgment distributor AD_i with L inputs and N outputs. In this case, the packet filters perform step 1 of Figure 4, whereas the prioritized concentrator performs step 2.

As the in the Knockout switch, the packet filters PF'_i accept packets destined for that output and ignore the rest. However, as compared to the known packet filter of the Knockout switch, each packet filter PF'_i has some additional features. Firstly, a packet filter prevents a packet from a corrupted frame from entering the concentrator. Secondly, a packet filter marks the first packet of frames as low priority packets and other packets as high priority packets. Figure 11 is a block diagram of an individual packet filter PF'_i, and Figure 12 is a flow chart illustrating the functions of an individual packet filter. An address filter 101 reads the destination address of the incoming packet and checks whether the packet is destined for the output in question (phase 110). If this is the case, the packet is forwarded to an identification unit 102, which identifies the packets which are the first packet of frames (phase 111) by reading the headers of the incoming packets. For example, if the packets are ATM cells, the third bit of the PTI field of the cell header indicates whether the cell is the last cell segmented from a larger data unit.

Packets which are not first packets are transferred for a further check to a marking unit 103 (phase 112). In this further check the acknowledgment information on the previous packet of the same connection is checked. If the previous packet has been accepted, the packet is marked with a high priority stamp (phase 114) and supplied to the output of the packet filter. Otherwise, the output of the filter is not enabled, i.e. the packet is prevented from entering the concentrator (the packet is discarded). Packets identified as first packets bypass the acknowledgment check and are directly marked with a low priority stamp before being supplied to the output of the packet filter.

The following table describes the operation of the packet filter in various situations.

Input	Acknowledgment	Priority status/discard	Output
first packet	1	low-priority	first-packet
first packet	0	low-priority	first packet
other packet	1	high-priority	other packet
other packet	0	discard	no packet
packet does not belong to the output, or no packet on the input	1	discard	no packet
packet does not belong to the output, or no packet on the input	0	discard	no packet
	1: previous packet has been accepted 0: previous packet has been discarded or no packet has been accepted		

In the second embodiment of the switch (Figure 10), the prioritized concentrator operates in the same way as in the first embodiment (Figure 5), i.e. it performs step 2 of Figure 4.

In the second embodiment, the acknowledgment distributors are distributed to every output module. Each distributor AD_i can be a simple L×N crossbar network with inputs connected to the outputs of the concentrator and outputs connected to the packet filters. If a packet passes through the concentrator PC_i , an acknowledgment is generated and sent back through the distributor to the packet filter from which the packet came. Figure 13 illustrates an example of the acknowledgment feedback path. The route of the packet is shown by a thick dashed line and the route of the acknowledgment by a thin dashed line.

Although the invention has been described here in connection with the examples shown in the attached figures, it is clear that the invention is not limited to these examples, as it can be varied in several ways within the limits set by the attached patent claims. For example, the principle described above

can also be applied to the individual switching elements of a multistage switching fabric. Depending on the basic structure of the switch to which the invention is applied, the switch may need a separate means for counting the number of simultaneously contending packets, i.e. the load level, so that the drop mechanism can be activated when the threshold is exceeded. However, a switch structure based on the knockout type switch is preferable in that no separate means are needed for measuring the load level. The switch can also have an additional drop mechanism which monitors the fill rates of the output buffers and drops packets from said buffers, i.e. the switch can have two superimposed drop mechanisms.

Patent claims:

1. A frame discard method for a packet switch, the method comprising the steps of

- receiving packets belonging to at least one transmission connection, consecutive packets of an individual transmission connection forming frames,

- switching packets from N input ports of the switch to N output ports of the switch, through at least one intermediate port,

- discarding complete frames, if the load level in the switch exceeds a predetermined threshold,

characterized in that

in a switch where the maximum number of packets that can be transmitted simultaneously to one port is smaller than N, the packets are discarded in such a way that

when the number of packets simultaneously contending for an individual port exceeds said maximum number, at least one first packet of a frame is chosen to be discarded from among said packets, and

once a first packet of a frame has been discarded, the remaining packets of the same frame are also discarded, regardless of the current buffering capacity of the switch.

2. A method according to claim 1, characterized in that at least one first packet of a frame is chosen randomly from among packets which are the first packet of a frame.

3. A method according to claim 1, characterized in that the other packets belonging to a frame whose first packet has been switched through the switch are discarded only if the switch has no buffering capacity for said packets.

4. A method according to claim 1, characterized in that the packets received are classified as high and low priority packets and supplied to a Knockout switch comprising Knockout concentrators in which packets compete pairwise in such a way that the high priority packet wins the low priority packet whenever high and low priority packets compete against each other.

5. A method according to claim 1, characterized in that said remaining packets are discarded at the input edge of the switch to prevent them from entering the switch.

6. A packet switch for switching packets, the switch comprising N input ports, N output ports, and at least one intermediate port between the input ports and the output ports, whereby the maximum number of packets that can be transmitted simultaneously to one port of the switch is smaller than N,

characterized in that

the switch further comprises

first means (IM_i, $PC_1...PC_N$; $PF'_1...PF'_N$, PC_i) for discarding at least one first packet of a frame when the number of packets simultaneously destined for an individual port exceeds said maximum number, and

second means (AD, IM_i; AD_i, $PF'_1...PF'_N$) for discarding the remaining packets of a frame, if the first packet of the same frame has been discarded.

7. A packet switch according to claim 6, characterized in that said first means comprise

classification means for dividing the packets received into low and high priority packets, and

Knockout concentrators ($PC_1...PC_N$; PC_i) comprising elements (SE) in which packets compete pairwise in such a way that a high priority packet always wins over a low priority packet when said packets compete against each other.

8. A packet switch according to claim 7, characterized in that it comprises

N parallel buses and N output modules, each output module being connected to each of the buses and including a Knockout concentrator, and

N input modules, each being connected to one of the input buses and each comprising said classifying means.

9. A packet switch according to claim 8, characterized in that the second means comprise a centralized distributor connected to the outputs of the Knockout concentrator of each output module for routing acknowledgments from each concentrator to the input modules, said acknowledgments containing information on packets having passed the concentrators.

10. A packet switch according to claim 9, characterized in that the centralized distributor is a crossbar matrix.

11. A packet switch according to claim 7, characterized in that the switch is a Knockout type switch comprising

N parallel buses and

N output modules, each comprising

N packet filters, each having an input and an output, the input of each filter being connected to one of the buses, and each filter comprising said classifying means (103),

one Knockout concentrator having N inputs and L outputs, each of the inputs being connected to the output of one of the packet filters, and

a distributor (AD_i) connected to the outputs of the concentrator for routing acknowledgments from the concentrator to the packet filters, said acknowledgments containing information on packets which have passed the concentrator.

12. A packet switch according to claim 11, characterized in that the distributor is a crossbar matrix.

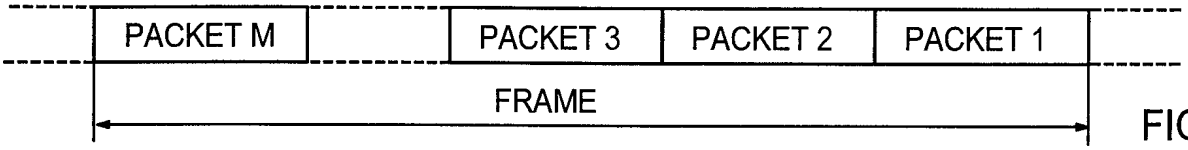


FIG. 1

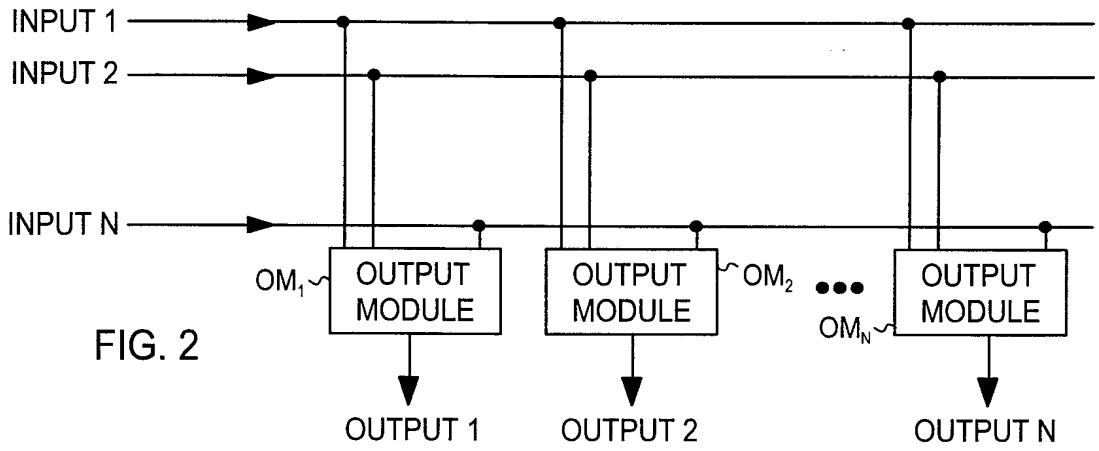


FIG. 2

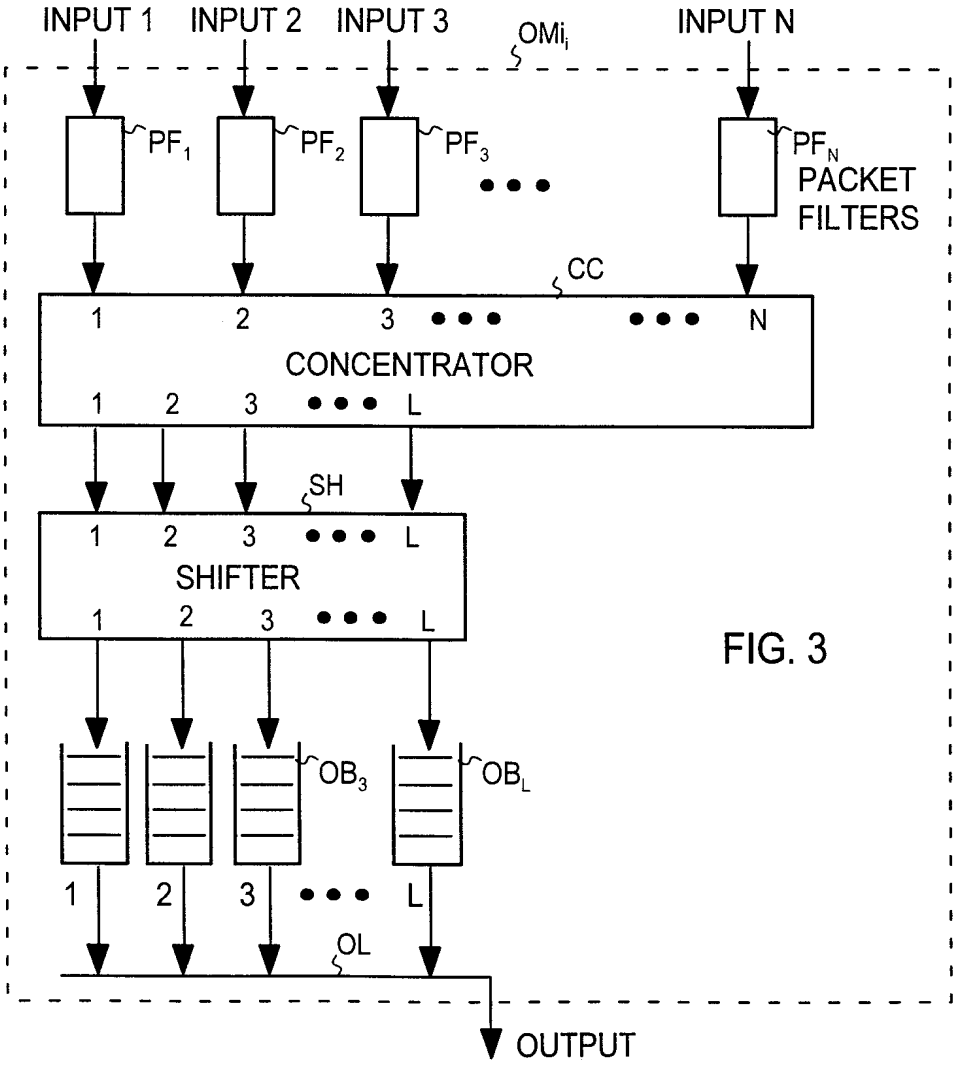


FIG. 3

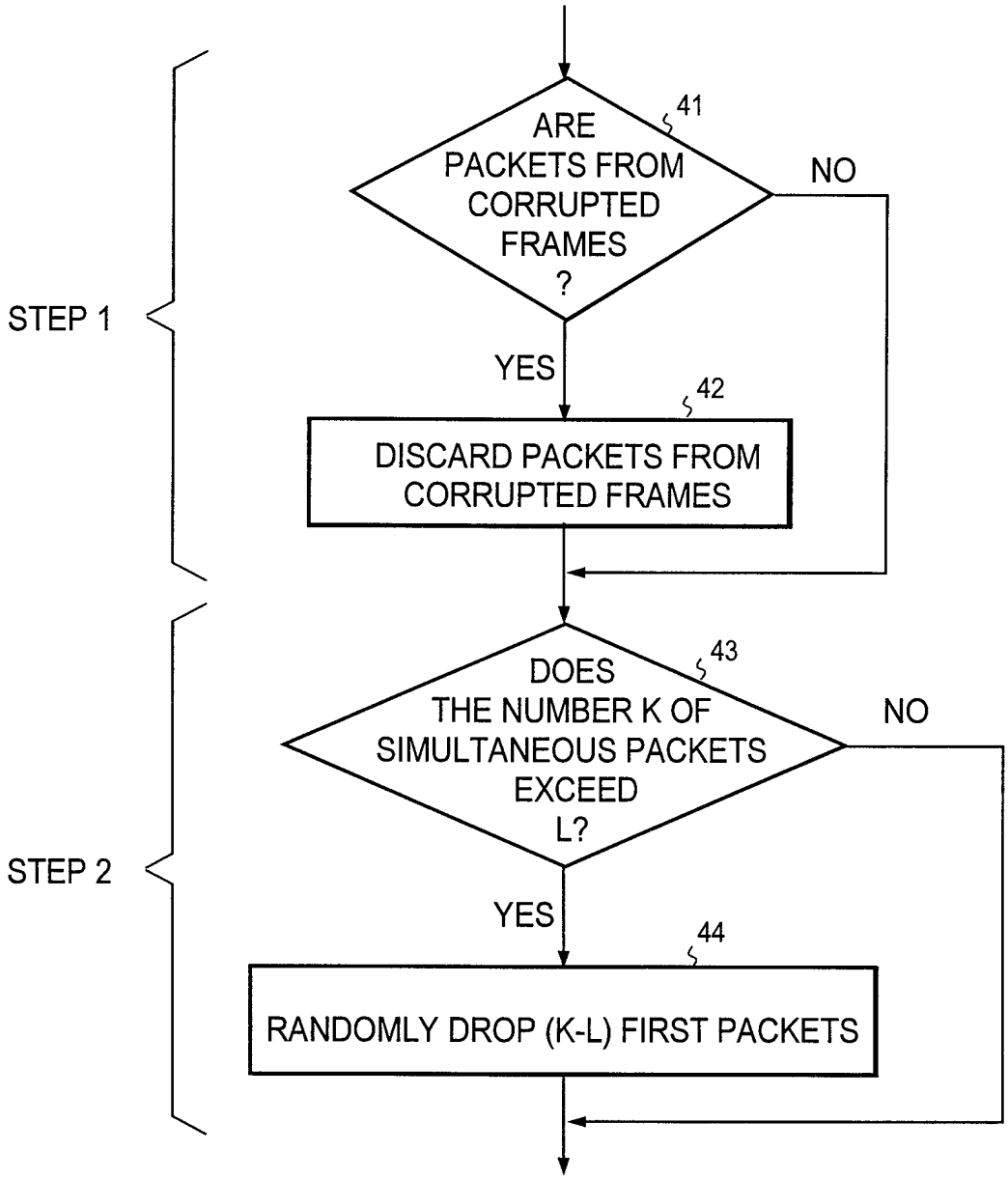

STEP 1
ARE PACKETS FROM CORRUPTED FRAMES ?
41
NO
YES
42
DISCARD PACKETS FROM CORRUPTED FRAMES
STEP 2
DOES THE NUMBER K OF SIMULTANEOUS PACKETS EXCEED L?
43
NO
YES
44
RANDOMLY DROP (K-L) FIRST PACKETS

FIG. 4

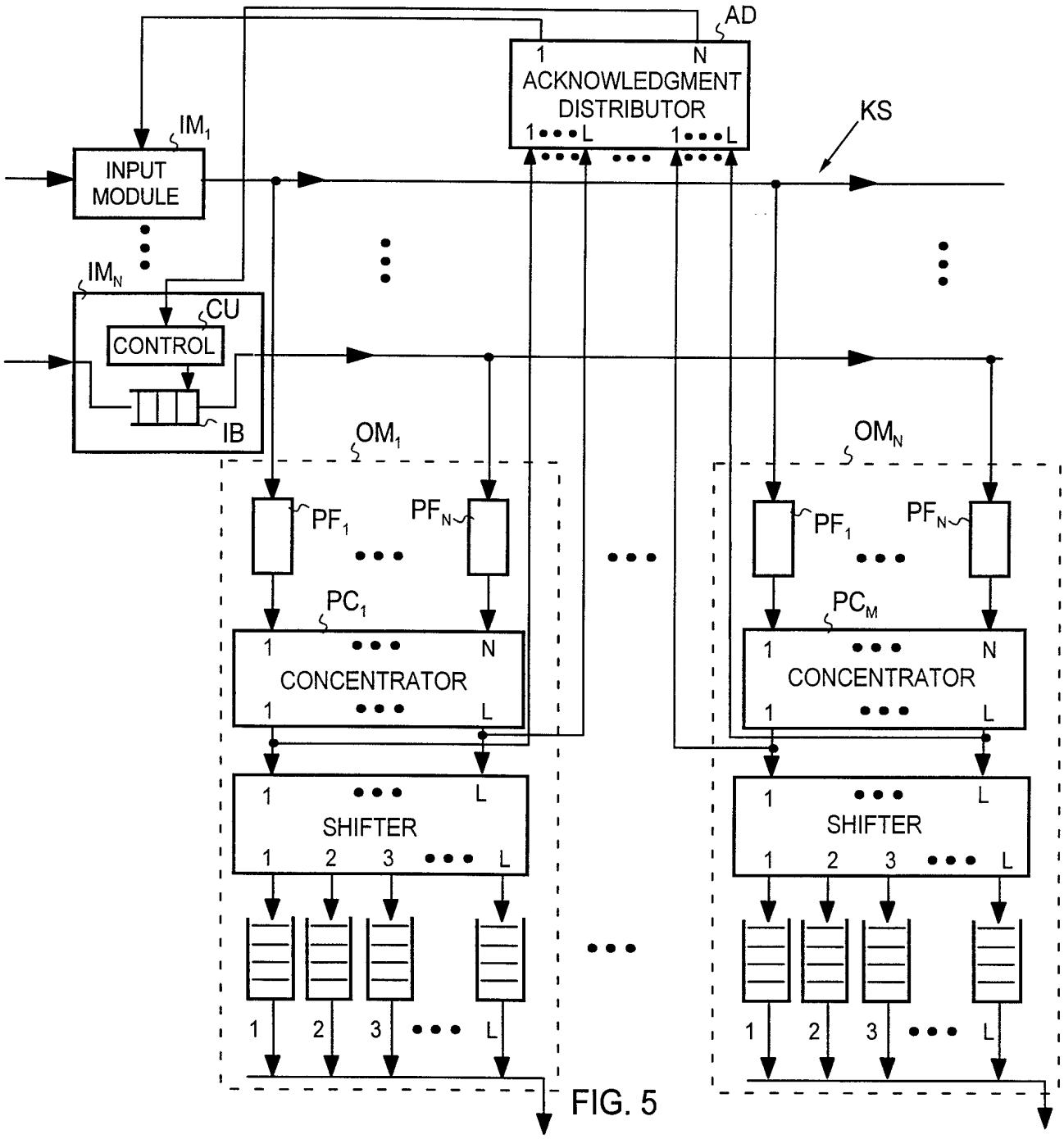


FIG. 5

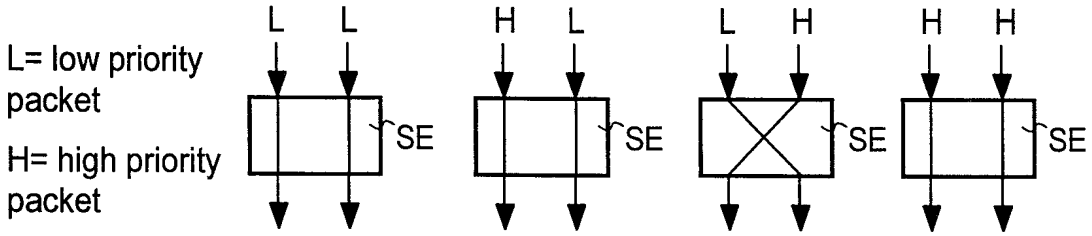


FIG. 6a FIG. 6b FIG. 6c FIG. 6d

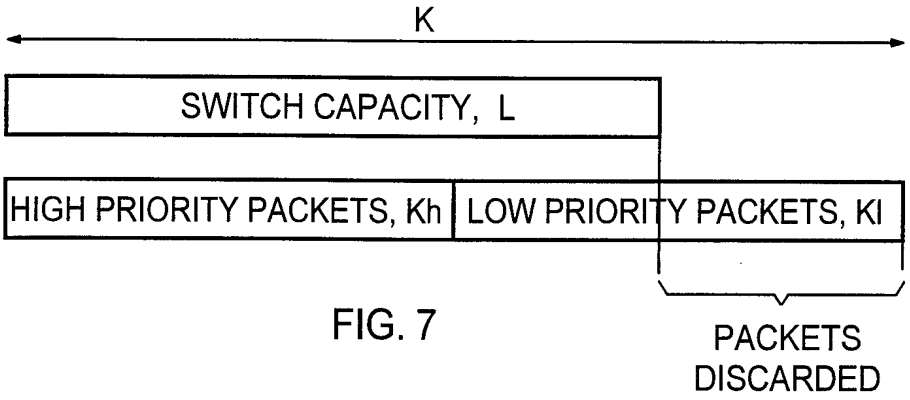


FIG. 7

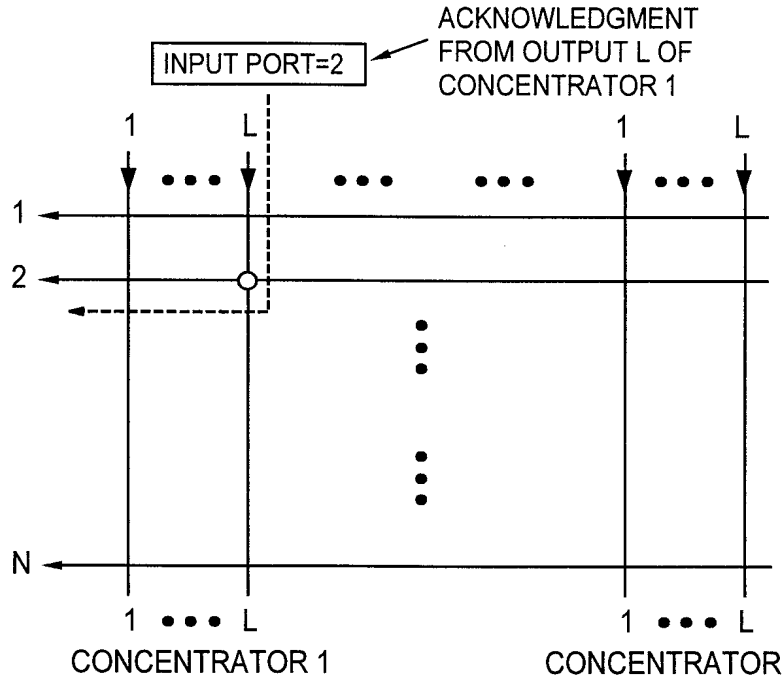


FIG. 8

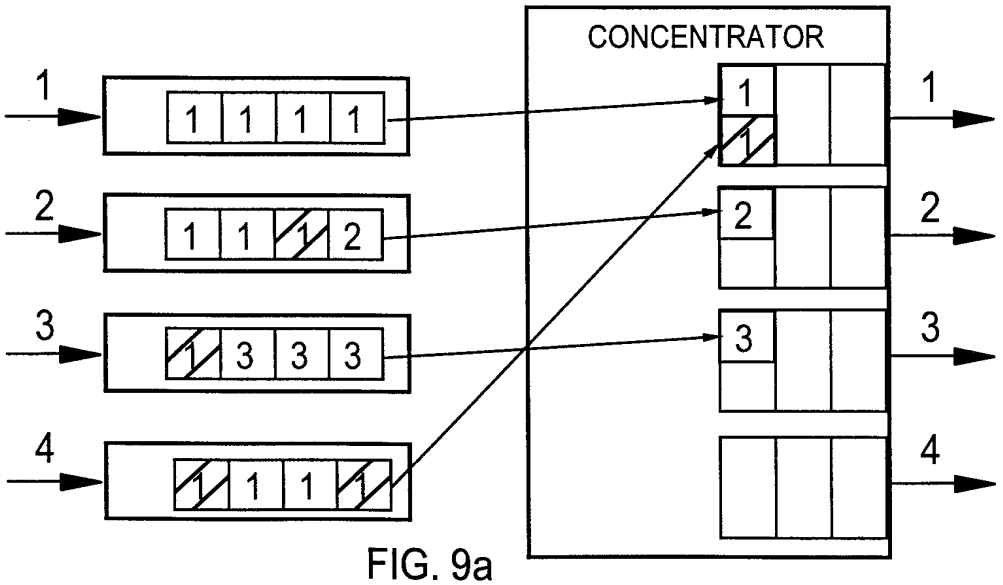


FIG. 9a

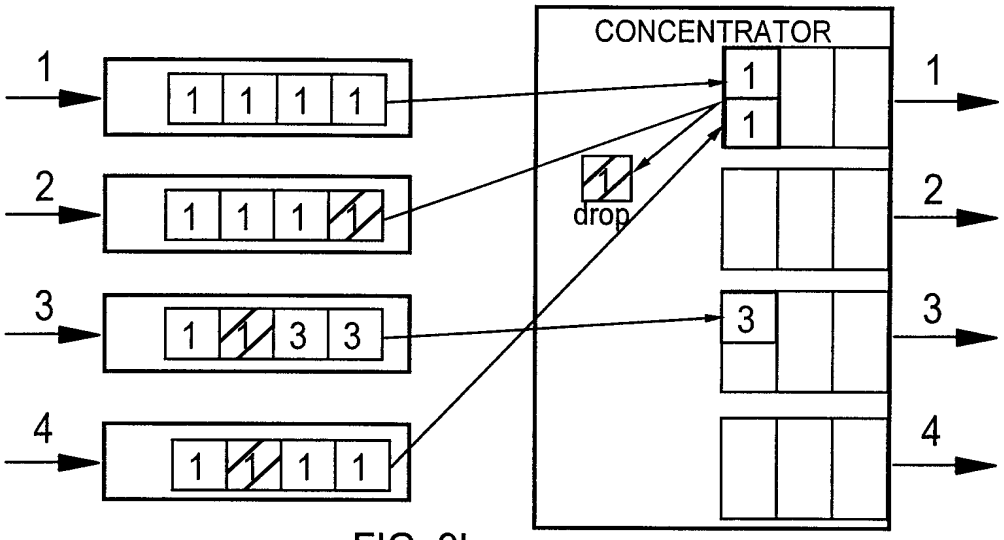


FIG. 9b

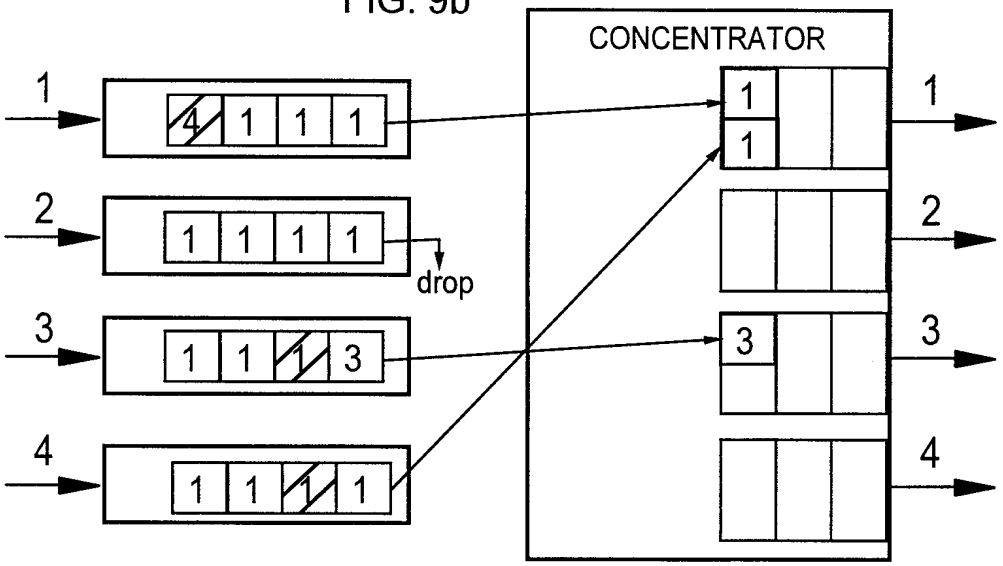


FIG. 9c

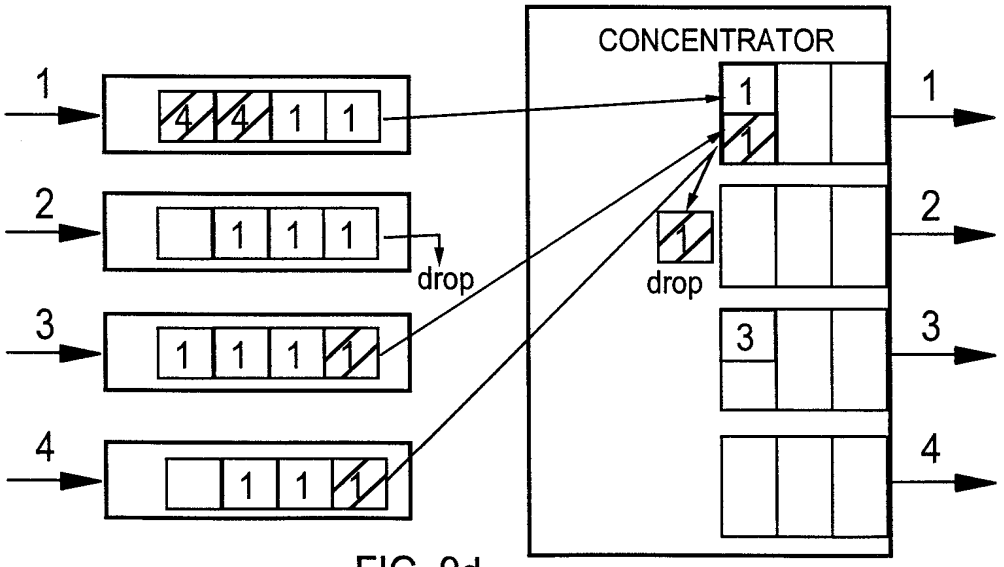


FIG. 9d

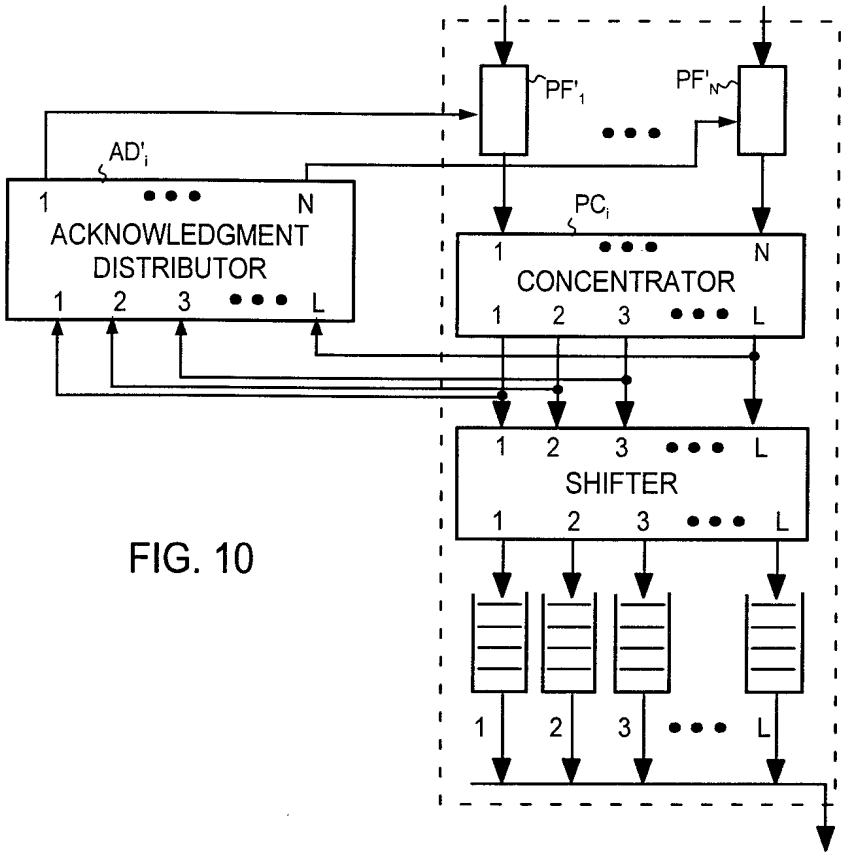

PF'1
PF'N
AD'i
1 ••• N
ACKNOWLEDGMENT
DISTRIBUTOR
1 2 3 ••• L
PCi
1 ••• N
CONCENTRATOR
1 2 3 ••• L
1 2 3 ••• L
SHIFTER
1 2 3 ••• L
1 2 3 ••• L

FIG. 10

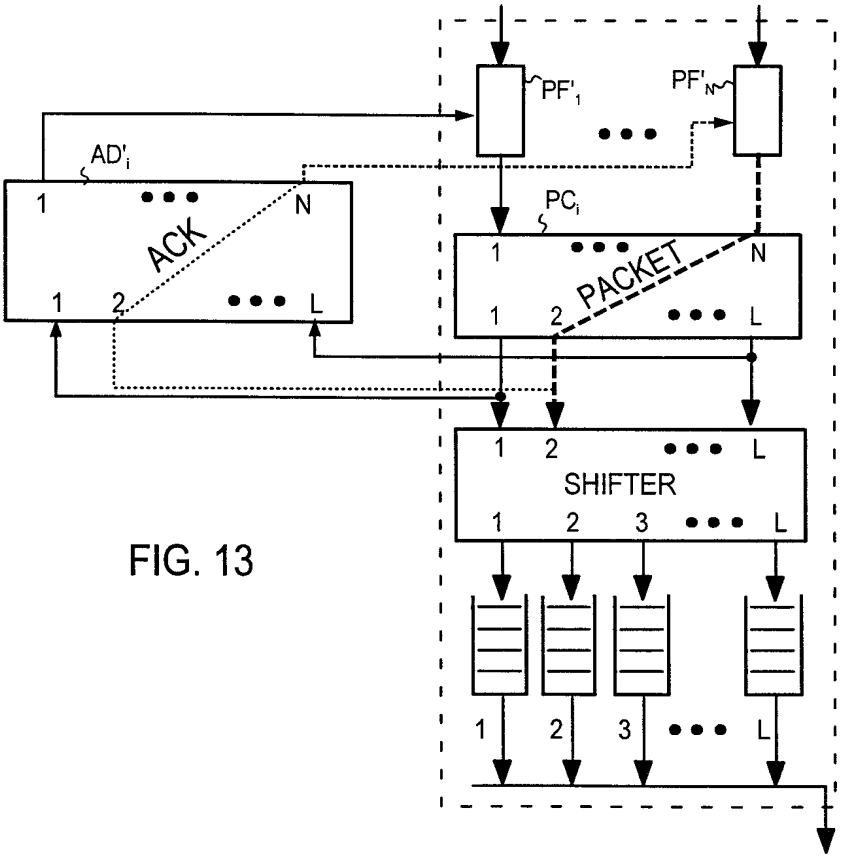

PF'1
PF'N
AD'i
1 ••• N
ACK
1 2 ••• L
PCi
1 ••• N
PACKET
1 2 ••• L
1 2 ••• L
SHIFTER
1 2 3 ••• L
1 2 3 ••• L

FIG. 13

FIG. 11

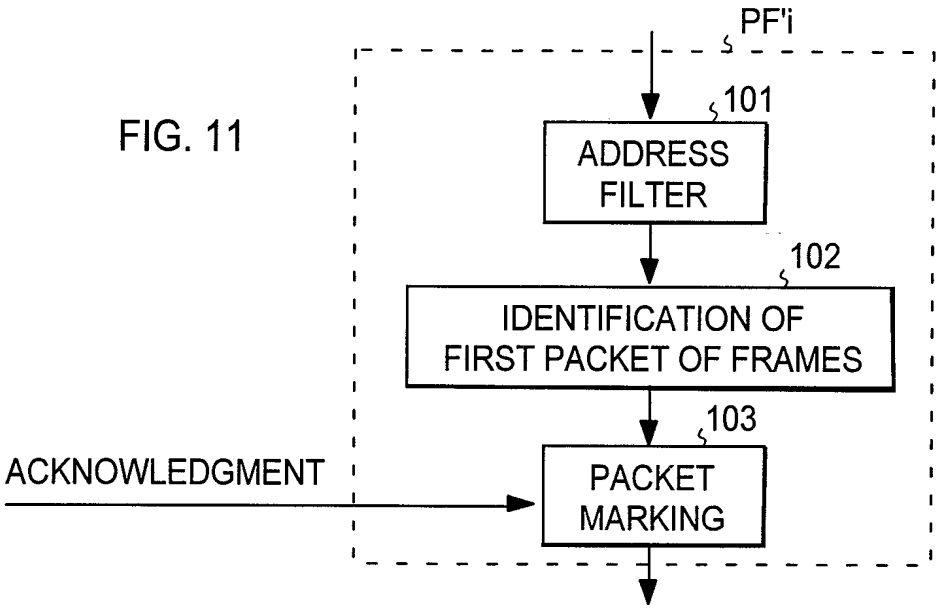
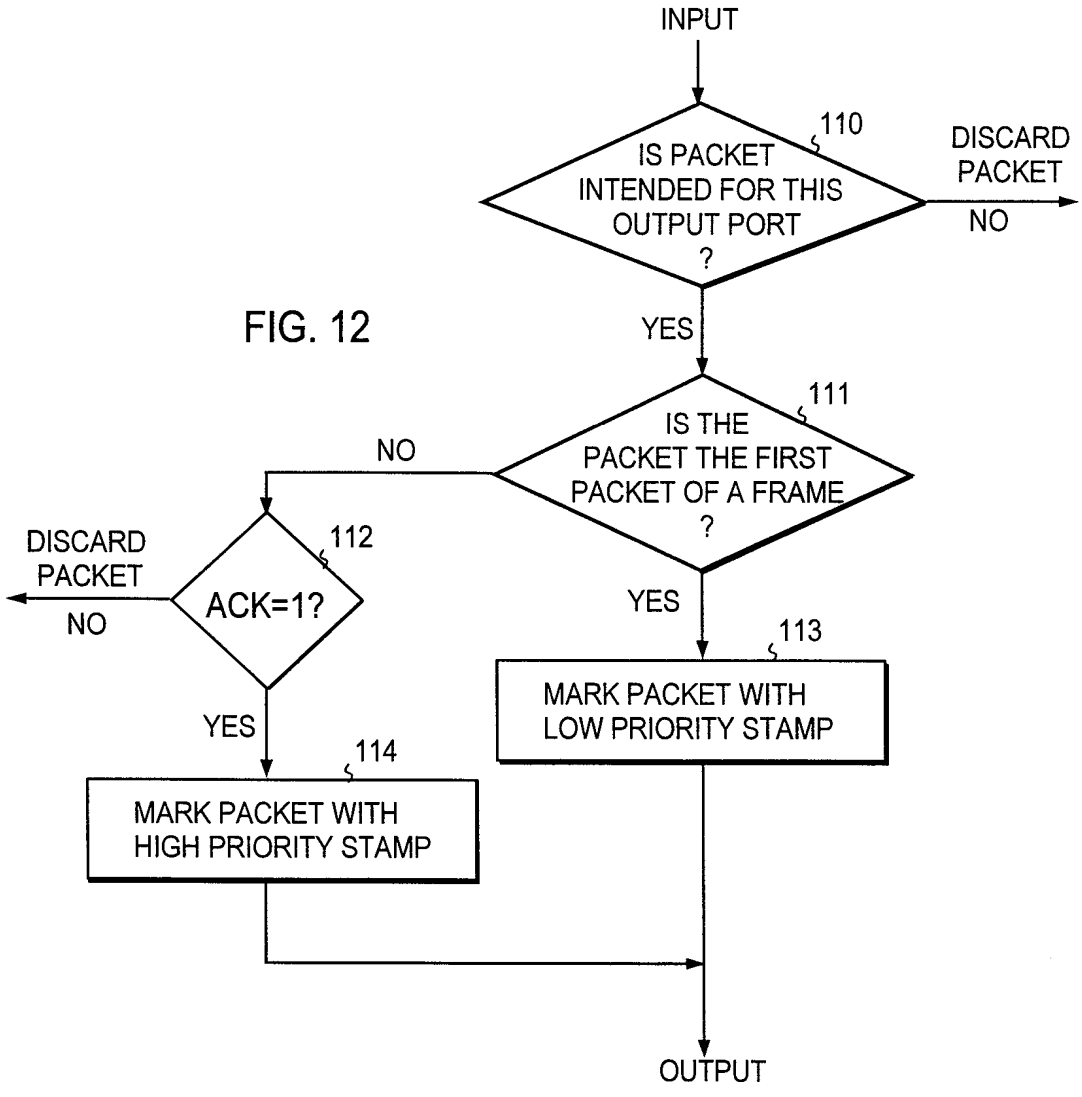

PF'i
101
ADDRESS FILTER
102
IDENTIFICATION OF FIRST PACKET OF FRAMES
103
PACKET MARKING
ACKNOWLEDGMENT

FIG. 12


INPUT
110
IS PACKET INTENDED FOR THIS OUTPUT PORT ?
DISCARD PACKET
NO
YES
111
IS THE PACKET THE FIRST PACKET OF A FRAME ?
NO
YES
112
ACK=1?
DISCARD PACKET
NO
YES
113
MARK PACKET WITH LOW PRIORITY STAMP
114
MARK PACKET WITH HIGH PRIORITY STAMP
OUTPUT

PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
H04Q 11/04, H04L 12/56

A3

(11) International Publication Number: **WO 99/25147**

(43) International Publication Date: 20 May 1999 (20.05.99)

(21) International Application Number: PCT/FI98/00872

(22) International Filing Date: 10 November 1998 (10.11.98)

(30) Priority Data:
974216 12 November 1997 (12.11.97) FI

(71) Applicant *(for all designated States except US)*: NOKIA TELECOMMUNICATIONS OY [FI/FI]; Keilalahdentie 4, FIN–02150 Espoo (FI).

(72) Inventor; and
(75) Inventor/Applicant *(for US only)*: MA, Jian [FI/FI]; Pihlajamarjantie 13–15 C1, FIN–02940 Espoo (FI).

(74) Agent: PATENT AGENCY COMPATENT LTD.; Teollisuuskatu 33, P.O. Box 156, FIN–00511 Helsinki (FI).

(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

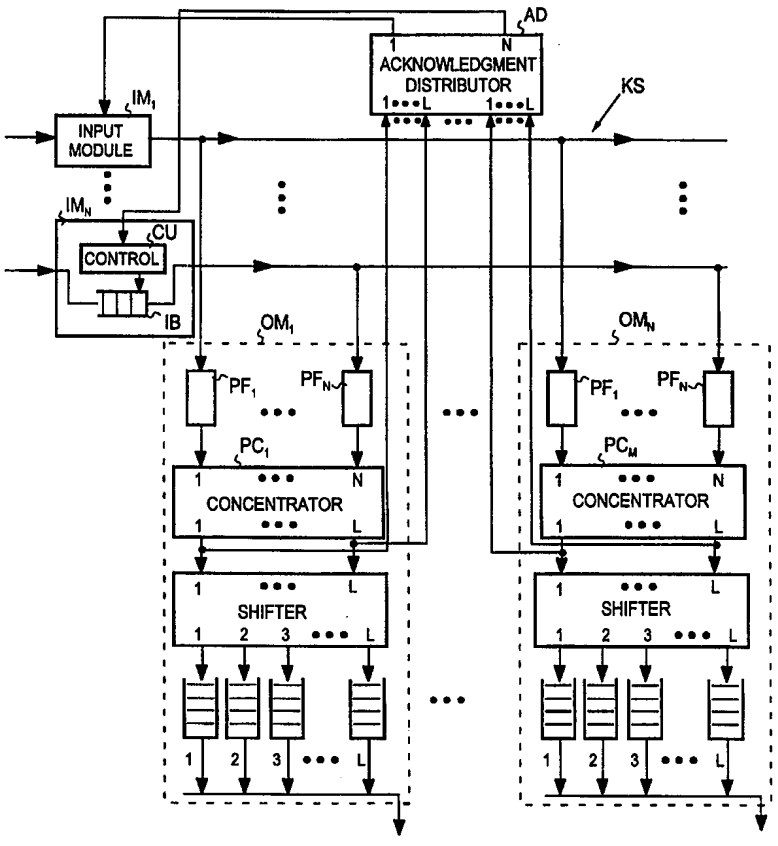
Published
With international search report.

(88) Date of publication of the international search report:
5 August 1999 (05.08.99)

(54) Title: A FRAME DISCARD MECHANISM FOR PACKET SWITCHES

(57) Abstract

The invention relates to a frame discard method for a packet switch and to a packet switch. The method comprises the steps of (a) receiving packets belonging to at least one transmission connection, consecutive packets of an individual transmission connection forming frames, (b) switching packets from N input ports of the switch to N output ports of the switch, through at least one intermediate port, and (c) discarding complete frames, if the load level in the switch exceeds a predetermined threshold. In order to combine a good frame level performance with implementation simplicity, the packets are discarded in a switch where the maximum number of packets that can be transmitted simultaneously to one port is smaller than N, in such a way that (i) when the number of packets simultaneously contending for an individual port exceeds said maximum number, at least one first packet of a frame is chosen to be discarded from among said packets, and (ii) once a first packet of a frame has been discarded, the remaining packets of the same frame are discarded, regardless of the current buffering capacity of the switch.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

INTERNATIONAL SEARCH REPORT

International application No.
PCT/FI 98/00872

A. CLASSIFICATION OF SUBJECT MATTER

IPC6: H04Q 11/04, H04L 12/56
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC6: H04Q, H04L, H04M

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

SE,DK,FI,NO classes as above

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EDOC, JAPIO, WPI, INTERNET

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 4760570 A (ANTHONY ACAMPORA ET AL), 26 July 1988 (26.07.88), column 2, line 18 - column 3, line 4, figures 1,2,5,6	1-12
Y	IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, Volume 13, No 4, May 1995, Allyn Romanow, "Dynamics of TCP Traffic over ATM Networks", page 636, column 3, paragraph 2, line 1 - page 637, column 1, line 7	1-12
A	US 4754451 A (KAI Y. ENG ET AL), 28 June 1988 (28.06.88), see the whole document	1-12

[X] Further documents are listed in the continuation of Box C. [X] See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" erlier document but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance: the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance: the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
25 May 1999	29 -05- 1999
Name and mailing address of the ISA/ Swedish Patent Office Box 5055, S-102 42 STOCKHOLM Facsimile No. +46 8 666 02 86	Authorized officer Erik Johannesson/mj Telephone No. +46 8 782 25 00

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/FI 98/00872

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
P,A	US 5764641 A (ARTHUR LIN), 9 June 1998 (09.06.98), column 2, line 59 - column 3, line 31, figures 3A, 3B --	1-12
A	EP 0323835 A2 (NEC CORPORATION), 12 July 1989 (12.07.89), see the whole document -- --------	1-12

INTERNATIONAL SEARCH REPORT

Information on patent family members

03/05/99

International application No.
PCT/FI 98/00872

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 4760570 A	26/07/88	CA 1278848 A	08/01/91
		DE 3788019 D,T	03/03/94
		EP 0256702 A,B	24/02/88
		SE 0256702 T3	
		JP 63086938 A	18/04/88
US 4754451 A	28/06/88	CA 1274303 A	18/09/90
		EP 0257816 A	02/03/88
		JP 2549667 B	30/10/96
		JP 63043447 A	24/02/88
US 5764641 A	09/06/98	NONE	
EP 0323835 A2	12/07/89	CA 1317014 A	27/04/93
		JP 1177239 A	13/07/89
		US 4868813 A	19/09/89